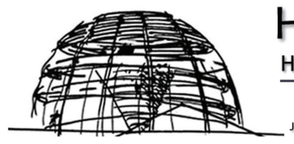
# Ultra-fast, accurate and cost-effective NGS read alignment with significant storage footprint reduction

**Bas Tolhuis, Jos Lunenberg & Hans Karten**

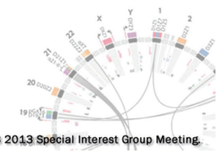
Genalice B.V. – Harderwijk, the Netherlands

For more information:
www.genalice.com
bas.tolhuis@genalice.com

GENALICE
TECHNOLOGY FOR PEOPLE & SCIENCE



HITSEQ 2013
High Throughput Sequencing
Algorithms & Applications
July 19-20, 2013 - ICC, Berlin, Germany - An ISMB/ECCB 2013 Special Interest Group Meeting



## Introduction

As the rate of next-generation sequencing increases, greater throughput is demanded from read aligners. The ideal read aligner needs to be capable of more than just simple alignment. It needs to:

1. Be fast, sensitive and accurate
2. Find longer, gapped alignments
3. Accept higher upper read lengths (be compatible with Roche/454 and Ion Torrent)
4. Be suitable for paired-end and split-end data
5. Run on commodity hardware

The full-text minute index is often used to make alignment as fast as possible and memory-efficient. The most widely used full-text minute index read aligners are Bowtie [1] and Burrows-Wheeler Alignment (BWA) [2].
Here we present the data of an innovative ultra-fast new read aligner that uses a novel algorithm and reduces the storage footprint of the output file in comparison to BWA.

## Methods

### GENALICE MAP

GENALICE developed an innovative next-generation sequence read alignment tool. The tool uses FASTQ format as input. It comprehensively combines read alignment and variant calling. Its output is flexible and can be our unique data footprint reducing format (GAR = GENALICE Aligned Reads), SAM/BAM output format or Variant Call Format (VCF).

### Data description

Data is derived from 1,000 Genomes Project and includes reads from human chromosome 20 of sample NA12878. It contains 2x23.6 million paired-end reads with a length of 101 bases per read. This results in an average coverage of 64 times chromosome 20. This data set has been recommended by Broad Institute to evaluate the GATK pipeline. As such this data set has been described in large detail.

### System configuration

We compared BWA and GENALICE MAP performance using the following hardware configurations. Our system has 2x Intel Xeon E5 2620 CPUs with 6 cores per CPU and 2 threads per core (24 cores in total). RAM memory is 96 GB and running a Linux x86-64 (open SUSE 12.2) operating system.

### Reference index build

Both BWA and GENALICE MAP require that the reference genome is indexed. We used GRCh37 release of the human genome as a reference. BWA builds its index in 1 hour, 12 minutes and 46 seconds using the hardware configuration described above. GENALICE MAP builds its index in 27 minutes and 44 seconds.

### Alignment

FASTQ files containing paired-end reads from chromosome 20 for NA12878 were aligned using either BWA (version 0.6.2-r126) or GENALICE MAP. Both tools were run on the hardware described above. BWA was run using 24 threads with default parameters.

## Results & Discussion

### Fast alignment

GENALICE MAP aligns reads faster than BWA. The runtime of BWA on 64x coverage of chromosome 20 is approximately 72 minutes. GENALICE MAP aligns that same data set in 47 seconds. This process includes preparation of the alignment in which the reference genome index is prepared. This preparation step takes 3 seconds. The second step is the alignment procedure itself and, finally, the aligned reads are written to disk (~ 5 seconds). The major part of runtime includes the alignment procedure (39 seconds).

The average alignment speed of GENALICE MAP is around 92 million bases per second, whereas BWA aligns with ~1.1 million bases per second (Figure 1).

### Alignment result

The alignment results of BWA and GENALICE MAP are highly comparable. When we compare depth of coverage between BWA and GENALICE MAP aligned reads we notice a high degree of similarity (Figure 3). The total number of reads is quite similar (Figure 4).

### Storage footprint reduction

GENALICE MAP uses a novel format to report its aligned reads, namely GENALICE Aligned Reads (GAR). This format is results in a significant storage footprint reduction compared to the commonly used BAM format and the unaligned FASTQ format (Figure 2). The GAR format is fully realignable and as such can replace both BAM and FASTQ files as format to store.

## Results & Discussion *continued*

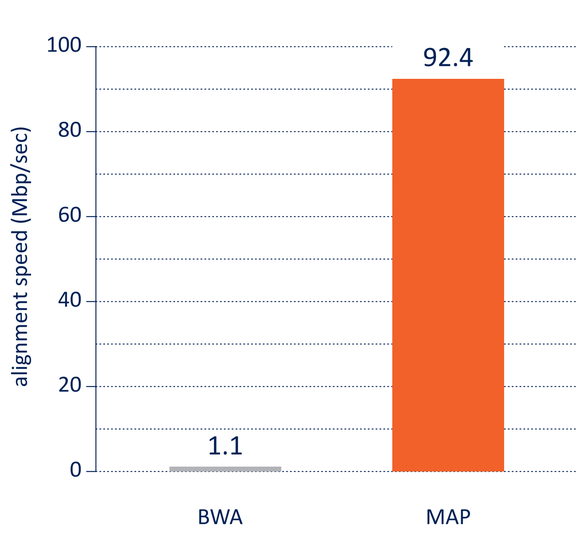



***Figure 1. Average alignment speed.***
*Performance of BWA (dark gray) and GENALICE MAP (orange) alignment speeds (megabases per second) are shown. For this data set BWA aligns at an average speed of 1.1 million bases per second. The average performance of GENALICE MAP is approximately 92 million bases per second.*

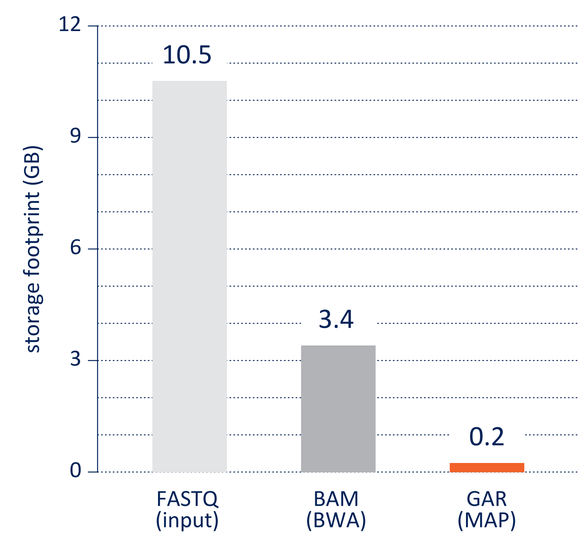



***Figure 2. Storage footprint NGS data formats.***
*NGS data formats from NA12878 (1,000 Genomes Project) chromosome 20 plotted as a function of the disk storage footprint in Gigabytes (GB). Three formats are compared. Input sequence reads in FASTQ format (light gray) require 10.5 GB disk space. BWA aligned reads into BAM format (dark gray) use 3.4 GB. GENALICE MAP's novel GAR format (orange) reduces disk space of all reads to 0.2 GB.*

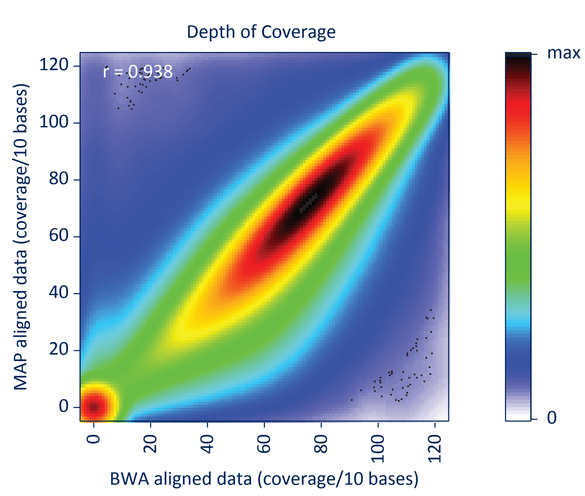



***Figure 3. Depth of coverage analysis.***
*Chromosome 20 was divided into consecutive bins that are 10 bases long. For each bin the average read coverage was calculated. The scatter plot compares depth of coverage of BWA aligned reads versus GENALICE MAP aligned reads. Heat map colors show density of the bins with white indicating no bins and black is the maximum density of bins. Pearson's correlation coefficient (r) is shown.*

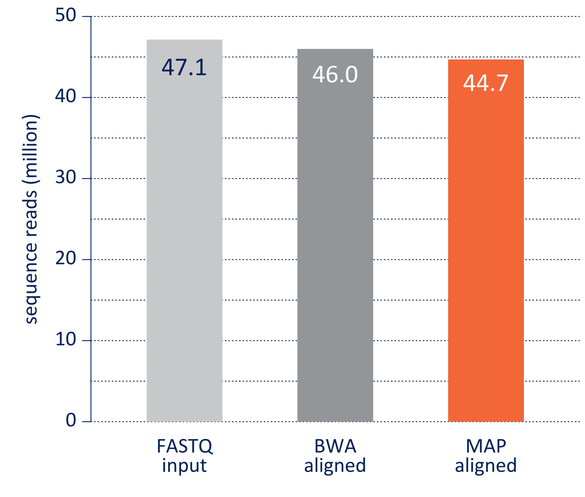



***Figure 4. Read number comparison.***
*Total read numbers plotted for input and aligned data. The input (light gray) is 2x23.6 million paired-end reads (47.1 million in total). BWA (dark gray) aligns 46.0 million reads with mapping quality larger than 0. GENALICE MAP (orange) aligns 44.7 million reads.*

### Discussion

GENALICE MAP greatly reduces the computing power and processing time needed to align NGS short reads, while being highly accurate. Moreover, the GAR format minimizes data storage capacity and facilitates data sharing. We think that GENALICE MAP is a cost effective alternative for existing open source read alignment tools due to its high speed, limited computing power usage and strongly reduced data storage footprint.

## Conclusion

- **Ultra-fast:** up to 200-fold faster than existing tools
- **High precision:** high accuracy in mapping reads (incl. long INDELs)
- **Small storage footprint:** 4GB for a complete human genome
- **Comprehensive:** combines alignment and variant calling in one tool
- **Flexible:** mapping accuracy independent of read length

**References**

1. Langmead B., Trapnell C., Pop M., and Salzber S., Ultra-fast and memory-efficient alignment of short DNA sequences to the human genome, (2009) *Genome Biology* 10:R25
2. Li H. and Durbin R., Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform, (2009) *Bioinformatics* 25:1754–1760